ResXplorer: Interactive Search for Relationships in Research Repositories

Laurens De Vocht[1], Selver Softic[2],
Erik Mannens[1], Rik Van de Walle[1], and Martin Ebner[2]

[1] Ghent University - iMinds, Multimedialab
Sint-Pietersnieuwstraat 41, 9000 Ghent, Belgium
{laurens.devocht,erik.mannens,rik.vandewalle}@ugent.be
[2] Graz University of Technology, IICM - Institute for Information Systems and Computer Media
Inffeldgasse 16c, 8010 Graz, Austria
{selver.softic,martin.ebner}@tugraz.at

Abstract. Research information is widely available on the Web. Peer-reviewed research publications as well related meta data from bibliography archives offers a vast of information to investigate about related publications, events and persons for a researcher. Usually the platforms supporting this information exchange have an API that allows access to the structured content or the information is already present as Linked Data. This opens a new way to search and explore research information. With the user interface of "ResXplorer" we help researchers to get an overview by using an approach that visualizes interactively search process in an aligned linked data knowledge base of research related resources. We show that visualizing resources, such as conferences, publications and proceedings, expose affinities between researchers and those resources. We characterize each affinity, between researchers and resources, by the amount of shared interests and other commonalities.

1 Introduction

Research 2.0 [1] as adaptation of the Web 2.0 for researchers defines researchers as main consumers of information. Typically researchers define queries with a set of keywords when searching for information related to their work, such as using Google[3] or digital archives like PubMed[4]. Linked Data delivers an entity based infrastructure to resolve the meanings of the keywords and the relations between them. We focus on the use of an interactive visualization to enable knowledge discovery using Linked Data. We aggregated in our Linked Data Knowledge Base datasets from DBLP(L3S)[5], COLINDA[6], DBPedia[7] and GeoNames[8].

[3] http://www.google.com
[4] http://pubmed.gov
[5] http://dblp.l3s.de/
[6] http://colinda.org
[7] http://dbpedia.org
[8] http://geonames.org

2 ResXplorer Architecture

ResXplorer[9] is an interactive interface for dealing with research datasets in the Linked Data cloud. It facilitates searching in data obtained from three open data repositories (DBpedia, COLINDA, DBLP), enriched with geographical information from GeoNames. A detailed architecture description is given in Figure 1.

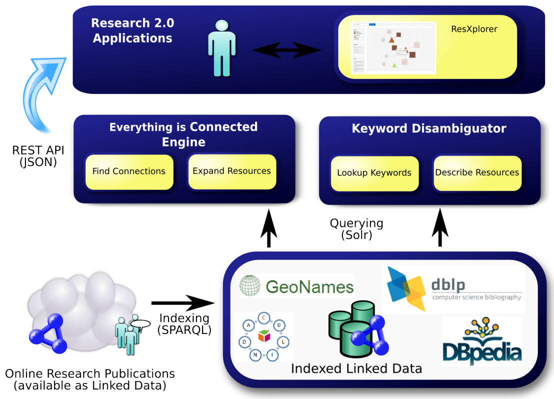

Fig. 1: In the ResXplorer architecture a Linked Data Index is used to dramatically increase search performance in the Knowledge Base.

2.1 Linked Data Knowledge Base

Our main domains of interest are scientific events, persons and publications. The DBLP[10] bibliography dataset as Linked Data is the base information source for our underlying search implementation. The COLINDA Linked Data set resolves scientific events their relatedness to bibliographical archives like DBLP. We used the *owl:sameAs* property to connect the conferences from COLINDA with corresponding proceedings instances from DBLP. Connecting COLINDA to GeoNames [2], as most important location related hub within Linked Data, uses the *swrc:location* property which points to the GeoNames location if they both correspond, while the connection of locations with DBPedia references was made by *dc:spatial* property.

2.2 Design Decisions

We chose a real-time keyword disambiguation to guide the researchers in expressing their research needs. We do this by allowing users to select the correct

[9] http://www.resxplorer.org
[10] http://dblp.l3s.de/dblp++.php

Table 1: Datasets used in the implementation.

Dataset	#Triples	#Instances	#Literals
DBPedia	332 089 989	27 127 750	161 710 008
DBLP (L3S)	95 263 081	13 173 372	17 564 126
COLINDA	143 535	15 788	70 334

meaning from a drop down menu that appears below the search box. Presenting candidate query expansion terms in real-time, as users type their queries, is useful during the early stages of the search [3]. In this is case it is very important that the users understand meaning of the suggested terms. Therefore we use an as straightforward as possible representation of the keyword mappings as shown in Figure 2.

Researchers enter the first keyword and select the matching mapping suggestion

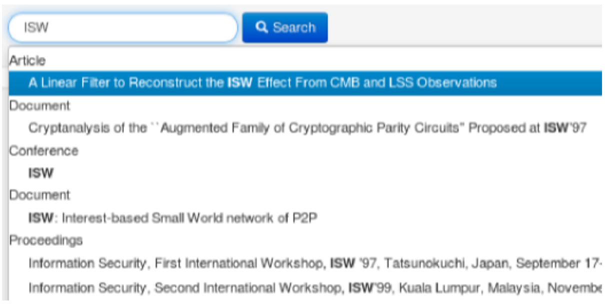

Fig. 2: Mapping of keywords to Linked Data entities.

that resolved their search focus at least one step forward. Parallel within choice and narrowing down the scope, the back-end connects the resources and ranks them according to the context. At the same time background modules also fetch neighbour links which match the selected suggestion.

Selection of various resources is then presented to the researchers. In case they have no idea which object or topic to investigate next, they get an overview of possible objects of interest (like points of interest on a street map) within radial interface.

A goal of the search is to explore information not seen before which makes it difficult to define an exact search goal. Besides allowing to search specific entities, our visualization facilitates exploratory browsing. This is particularly useful when information seeking with unclear defined search targets [4].

To enhance the guidance for users during search we use several visual aids of which the three most visible are:

1. **Shape:** We group sets of types in large groups and represent them using a shape. Types that cannot be assigned a group are grouped in a category 'Miscellaneous'.
2. **Color:** Every entity has a type and associated unique color. For a certain result set the user gets an immediate impression of the nature of the found resources.
3. **Size:** We rank each entity according to novelty and relation to the context and enlarge those that should attract attention from the researcher first.

Figure 3 shows how researchers can track the history of their search: the explored relations are marked red and clearly highlight the context of a search. Researchers can click on a list of resources they have searched to focus the visualization.

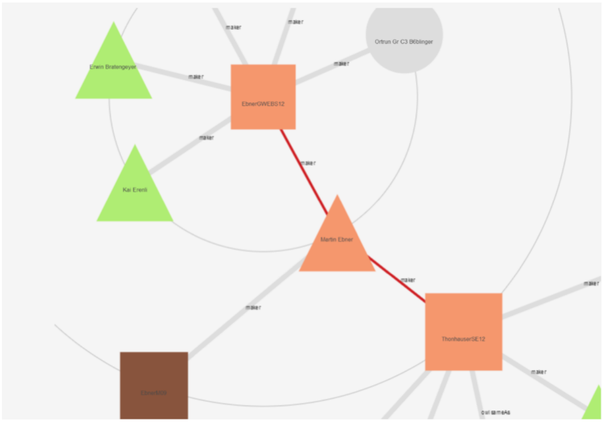

Fig. 3: A red line marks the explored relations in the visualized search context.

2.3 Implementation

We have developed the interface for search in a research linked data knowledge base combining the latest Linked Data technologies with an advanced indexing and path finding system. We built our implementation upon our earlier work using the "Everything is Connected" engine (EiCE) [5] EiCE and Web 2.0 technologies (such as JQuery and Django). The interface itself is a realization in HTML5 and Javascript making advanced use of JQuery UI in combination with the "Javascript Information Visualization Toolkit"[11].

[11] http://philogb.github.io/jit/

3 Visualizing Relations between Resources

We find relations between resources after matching the input given by the researcher in the knowledge base. With the delivery of first results, our engine expands the query and enhances the context. In the visualization we emphasize the affinities by showing, on a radial map [6], how the current focused entity is related to the other found entities. It is based on the concept of affinity that can be appropriately expressed in visual terms as a spatial relationship: proximity [7]. We additionally express the amount of unexpectedness as *novelty* of a resource in each particular search context. A typical example is illustrated in Figure 4.

Each time a combination of various resources is visualized, the application

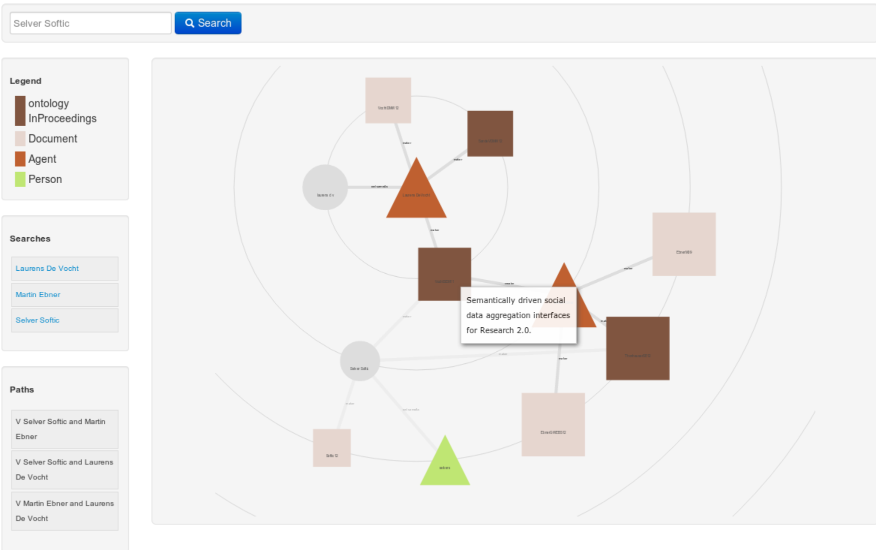

Fig. 4: Interface illustrating how *Laurens De Vocht* relates to *Selver Softic* based on common publications such as the highlighted "*Semantically...Research 2.0*".

suggests new queries: they are generally most useful for refining the system's representation of the researcher's need. In case they have no idea which entity to focus on or what topic to investigate next they get an overview of possible entities of interest, like points of interest on a street map. By profiling their activities and contributions on Social Media and other platforms such as their own research publications, the affinity with the proposed resources is enhanced. An iteration can consist of either one of two actions:

1. **Query Expansion:** The user expands the query space by clicking the results retrieved by initial keyword based search.The resolution of results hap-

pens based upon the properties of Linked Data like *rdf:label*, *owl:sameAs*, *rdf:seeAlso*, *dc:title* or *dc:description.*

2. **Additional Query Formulation:** Additional query expansion happens either through adding further keywords as well as through keyword combinations already entered where the back-end tries to deliver additional results based upon connection paths between the resources.

4 Usage

The result is a semantic search application providing both a technical demonstration and a visualization that could be applied in many other disciplines beyond Research 2.0. The main contribution of our work is, besides retrieving resources from Linked Data repositories, allowing researches to interactively explore relations between the resources and entities like events,publications or persons related to their work.

5 Acknowledgement

The research activities that have been described in this paper were funded by Ghent University, iMinds (Interdisciplinary institute for Technology) a research institute founded by the Flemish Government, Graz University of Technology, the Institute for the Promotion of Innovation by Science and Technology in Flanders (IWT), the Fund for Scientific Research-Flanders (FWO-Flanders), and the European Union.

References

1. Ullmann, T.D., Wild, F., Scott, P., Duval, E., Vandeputte, B., Parra Chico, G.A., Reinhardt, W., Heinze, N., Kraker, P., Fessl, A., Lindstaedt, S., Nagel, T., Gillet, D.: Components of a research 2.0 infrastructure. In: Lecture Notes in Computer Science,, Springer (2010) 590–595
2. Volz, R., Zhou, L., Finin, T., Joshi, A.: Towards ontology-based disambiguation of geographical identifiers. In: I3: Identity, Identifiers, Identification Workshop, World Wide Web Conference (WWW 2007). (2007) 8–12
3. White, R.W., Marchionini, G.: Examining the effectiveness of real-time query expansion. Information Processing & Management **43**(3) (2007) 685–704
4. Pace, S.: A grounded theory of the flow experiences of web users. International journal of human-computer studies **60**(3) (2004) 327–363
5. De Vocht, L., Coppens, S., Verborgh, R., Vander Sande, M., Mannens, E., Van de Walle, R.: Discovering meaningful connections between resources in the web of data. In: Proceedings of the 6th Workshop on Linked Data on the Web (LDOW). (2013)
6. Yee, K.P., Fisher, D., Dhamija, R., Hearst, M.: Animated exploration of dynamic graphs with radial layout. In: Proceedings of the IEEE Symposium on Information Visualization (INFOVIS). (2001)
7. Pintado, X.: The affinity browser. In: Object-oriented software composition. Prentice Hall (1995) 245–272

Appendix A

MINIMAL REQUIREMENTS
1. The application has to be an end-user application, i.e. an application that provides a practical value to general Web users or, if this is not the case, at least to domain experts.
✓*ResXplorer is an end-user application with a search and visual exploration interface that targets a special group of users: researchers. It enables them to find related persons, events and publications.*
2. The information sources used: should be under diverse ownership or control, should be heterogeneous (syntactically, structurally, and semantically), and should contain substantial quantities of real world data (i.e. not toy examples).
✓*As described in the architecture section about the underlying Linked Data Knowledge base, the application uses heterogenous data sources from four different Linked Data sets (DBLP, COLINDA, DBPedia, GeoNames)*
3. The meaning of data has to play a central role. Meaning must be represented using Semantic Web technologies. Data must be manipulated and processed in interesting ways to derive useful information and this semantic information processing has to play a central role in achieving things that alternative technologies cannot do as well, or at all.
✓*The novel approach about ResXplorer is incorporated within the application of regular search combined with specific link path finding module from the EiCE engine. Further enhancement of traditional keyword-based typeahead search interface with radial exploration screen and disambiguation of entities in different shapes, colors and sizes.*

ADDITIONAL DESIRABLE FEATURES
1. The application provides an attractive and functional Web interface (for human users).
✓*Out of the perspective of interactive search with radial mapping of resources and context centered interaction possibilities for query expansion, this feature can be considered as fulfilled.*
2.The application should be scalable (in terms of the amount of data used and in terms of distributed components working together). Ideally, the application should use a curated selection of published datasets of any size.
✓*The setup of our Linked Data Knowledge Base is only exemplary and a use case chosen for the this challenge. However extension of the functionality for other data sources from the Linked Open Data Cloud is easily possible by indexing additional data sets.*
3. Rigorous evaluations have taken place that demonstrate the benefits of semantic technologies, or validate the results obtained.
✓*We ran a set of 10 different queries and evaluated mean average precision of the system with respect to the extension of results through path finding. First evaluation delivered encouraging result of a mean average precision in the range of 60%.*

4. Novelty, in applying semantic technologies to a domain or a task that have not been considered before.

✓*The novelty lies on retrieval of results which is enriched based upon the structure of Linked Data and hiding the complexity of the retrieval model using only two interaction types: expanding and keyword querying.*

5. Functionality is different from or goes beyond pure information retrieval.

✓*Beyond merely retrieving resources and their context, we allow researchers to discover how two resources are related to each other in a specific context. Additionally by manual expanding via the radial interface facilitates searching in the results connected directly to the expanded node.*

6. The application has clear commercial potential and/or large existing user base.

✓*The biggest commercial potential lies within the underlying EiCE engine which enables efficient expanding the search context with high precision values by finding the paths within the retrieved results. Our implementation, including the indexing, is independent of the used Linked Datasets.*

7. Contextual information is used for ratings or rankings.

✓*The path finding and query expansion module of EiCE engine are using the context information to rerank and outline the importance of results discovered as described in [5].*

8. Multimedia documents are used in some way.

✗*Since we resolve also Linked Data from DBPedia, it would be possible to include the media data like e.g. images from the flickr wrapper. However current state of the implementation does not support media out of the box.*

9. There is a use of dynamic data (e.g. workflows), perhaps in combination with static information.

✓*We can integrate new Linked Data as it becomes available: the interface can assume a dynamically changing DBLP, for example. We can also handle changing models and do not need to adapt the interface to support new data models. Static information could be the ontologies used to model the data, based on this information we can enrich the visualization.*

10. The results should be as accurate as possible (e.g. use a ranking of results according to context).

✓*Inner ranking is based on the search preferences of single user as well on the distances of resolved resources to the point of interest specified as focus of search (the centered resource in the radial interface) and uses the affinity based ranking, which will be extended in future work to the researcher profiles based on Social Media.*

11. There is support for multiple languages and accessibility on a range of devices.

✓✗*This feature is not supported by now (currenly only English data sets indexed), however our infrastructure is extensible in this direction.*